Assignment 2: Operational Classification Report

Team 11

1. Operational Decision Rule

The operational goal is to minimize **False Positives (FP)** due to critically scarce resources (lifeboat capacity), prioritizing maximum **Precision**. The chosen operating threshold is $\tau = \mathbf{0.6}$.

2. Model Evidence: Performance Uplift and Feature Engineering

The engineered model, which used the BigQuery ML `TRANSFORM` clause to create features like family_size, fare_bucket, and the critical sex_pclass interaction, significantly improved predictive power over the canonical baseline. This improvement is attributed to the sex_pclass term, which directly models the historical ”women and children first” policy biases present in the survival data.

2.1. Baseline vs. Engineered Model Comparison

The table below demonstrates the performance lift gained from the feature engineering effort [2]. The increase in ROC AUC confirms better discriminative ability.

Table 1: Baseline vs. Engineered Model Performance Comparison

Metric	**Baseline Model (Simple Features)**	**Engineered Model (TRANSFORM)**	**Improvement**
ROC AUC	~ 0.83	$\sim \mathbf{0.85}$	**+2.0 pp**
Accuracy (@0.5)	~ 0.80	$\sim \mathbf{0.82}$	**+2.0 pp**
Log Loss	~ 0.5132	$\sim \mathbf{0.3678}$	Lower/Better (Better Calibration)

2.2. Confusion Matrix at Operating Threshold ($\tau = \mathbf{0.6}$)

The custom threshold of $\mathbf{0.6}$ was chosen to deliberately reduce False Positives (FP) at the expense of higher False Negatives (FN), maximizing resource allocation efficiency (Precision).

Table 2: Confusion Matrix at Proposed Operating Threshold ($\tau = \mathbf{0.6}$)

Predicted	**Actual Non-Survivor (0)**	**Actual Survivor (1)**
Predict Non-Survivor (0)	True Negatives (**TN**): 94	False Negatives (**FN**): 35
Predict Survivor (1)	False Positives (**FP**): 8	True Positives (**TP**): 34

3. Deployment Policy and Cost Analysis

3.1. Expected-Cost Analysis (Cost Matrix)

The deployment is **Global**—the same model and threshold ($\tau = 0.6$) are applied universally to all passengers, as resource allocation applies equally. The cost matrix reflects a policy where failing to save a potential survivor (FN) has a higher penalty than misallocating a limited resource (FP).

Table 3: Hypothetical Expected Cost Matrix (Cost per Error)

Actual	**Predicted Non-Survivor (0)**	**Predicted Survivor (1)**
Non-Survivor (0)	True Negative Cost: $0	False Positive Cost: $\$\mathbf{1,000}$
Survivor (1)	False Negative Cost: $\$\mathbf{4,000}$	True Positive Cost: $0

The total expected cost at the $\tau = \mathbf{0.6}$ threshold is calculated as:

$$\text{Total Cost} = (\text{FP} \times \$1,000) + (\text{FN} \times \$4,000)$$

$$\text{Total Cost} = (8 \times \$1,000) + (35 \times \$4,000) = \$8,000 + \$140,000 = \$\mathbf{148,000}$$

3.2. Fairness Observation (Precision Parity)

We assessed fairness by ensuring the scarce resource (lifeboat spot/Precision) is allocated equitably across the Sex subgroup, based on a policy threshold of a **5** percentage point (pp) gap.

- **Subgroup: Female Precision:** 0.841
- **Subgroup: Male Precision:** 0.827
- **Absolute Parity Gap:** $\mathbf{1.4}$ pp (The gap is well below the 5 pp policy limit, confirming operational fairness.)

4. Continuous Monitoring Plan

A robust plan is required to detect **Data Drift** (changes in input data distribution) and **Concept Drift** (changes in the underlying survival relationship) over time.

Table 4: Continuous Model Monitoring Schedule

Metric	**Threshold/Alert Condition**	**Cadence**	**Purpose (What are we tracking?)**
Calibration Error (Log Loss)	↑ 0.45 (Increase)	Weekly	Tracks model confidence and statistical fit. A sudden rise signals **Data Drift**.
Precision Parity Gap (Sex)	↑ 5 pp (Exceeds policy limit)	Monthly	Ensures ethical and policy compliance by tracking bias. Signals **Fairness Degradation**.
Feature Contribution Drift	↑ 10% change in feature weight	Monthly	Detects changes in the influence of key features (e.g., fare_bucket) could signal **Concept Drift**.

References

References

[1] Unit 2 Project Rubric, Defining the scope for BQML classification exercise.

[2] Model Training and Feature Engineering Notebooks (e.g., *Unit2__Zijing_Zhang__BQML.ipynb*, *Unit2_EthanLouie_BQML.ipynb*).

[3] Policy and Threshold Analysis, *model_governance_report.tex* and *ops_brief.tex*.

[4] Titanic Dataset. Source: `https://www.kaggle.com/datasets/yasserh/titanic-dataset`